International Journal of Computing and Artificial Intelligence

E-ISSN: 2707-658X
P-ISSN: 2707-6571
Impact Factor (RJIF): 5.57
https://www.computersciencejournals.com/ijcai/
IJCAI 2026; 7(1): 73-86
Received: 24-10-2025
Accepted: 30-12-2025

Khushboo Pawar
LNCT University, Bhopal, Madhya Pradesh, India

Dr. Devdas Saraswat
LNCT University, Bhopal, Madhya Pradesh, India

Corresponding Author:
Khushboo Pawar
LNCT University, Bhopal, Madhya Pradesh, India

Hybrid big data time series forecasting of Wind Power Using ARIMA, LSTM, GRU, and CNN-LSTM with pattern-based sequence clustering

Khushboo Pawar and Devdas Saraswat

DOI: https://www.doi.org/10.33545/27076571.2026.v7.i1b.244

Abstract
This study addressed the challenge of accurate wind power forecasting by developing a hybrid big data time-series model that integrated ARIMA, LSTM, GRU, and CNN-LSTM architectures with pattern-based sequence clustering. Using a real-world dataset containing 34,080 time-stamped observations recorded at 15-minute intervals, the research utilized ten meteorological and operational variables, including wind speed, preliminary power output, wind direction, temperature, humidity, atmospheric pressure, and rounded turbine measurements, along with YD15, a 15-minute-ahead power target used for supervised learning. A comprehensive preprocessing workflow—comprising outlier removal, missing-value interpolation, normalization, and feature engineering—was applied to ensure data quality. Dynamic Time Warping (DTW) clustering was employed to group similar temporal sequences, enabling localized model training across diverse wind regimes. The hybrid architecture was deployed in a distributed environment using Apache Spark, ensuring scalability and high processing throughput. Experimental evaluation on the dataset demonstrated that the hybrid model consistently outperformed standalone approaches, achieving lower MAE and RMSE and higher Accuracy, Precision, Recall, and F1-scores. Overall, the study provided a robust, scalable, and data-driven forecasting solution capable of capturing both linear and nonlinear wind power dynamics, supporting more reliable smart-grid operations and sustainable energy management.

Keywords: Hybrid time-series forecasting, Wind power prediction, ARIMA-LSTM-GRU-CNN-LSTM, Dynamic Time Warping (DTW) clustering, Big data analytics, Apache Spark, Renewable energy forecasting, Short-term power prediction, Meteorological data modelling, Smart grid optimization

Introduction

Wind power has become one of the cleanest and most promising substitutes to conventional fossil fuel-based electricity generation. Its growing integration into national and regional power grids contributes in a major way to curbing carbon emissions and meeting international clean energy targets (Hanifi *et al*., 2020) [13]. Nevertheless, because of its intrinsic intermittency and randomness, the electricity generation from wind farms varies greatly over a period of time. Such variability presents serious issues in terms of grid stability, load balancing, and reserve unit commitment planning (Neshat *et al*., 2021) [23]. To offset such issues, robust wind power forecasting becomes a critical necessity for optimal power system operation, facilitating forward-thinking decision-making in energy scheduling and reserve determination.

Conventional time series forecasting techniques, including the use of the Autoregressive Integrated Moving Average (ARIMA) model, have been the primary methodologies to be applied in renewable energy forecasting applications (Mohapatra *et al*., 2023) [22]. ARIMA is particularly suitable for identifying linear relationships and short-order dependencies with interpretable model parameters and solid baseline predictions (Yang *et al*., 2022) [31]. Nevertheless, ARIMA and related traditional statistical models are not capable of capturing nonlinear, chaotic, and non-stationary patterns embedded in meteorological and wind data (Dhakal *et al*., 2022) [7]. Therefore, their predictive ability weakens under intricate temporal changes and long-order dependencies.

By contrast, deep learning architectures like Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRU), and Convolutional Neural Network-Long Short-Term

Memory (CNN-LSTM) have proven to have robust ability in learning nonlinear temporal features and long-range dependencies from large datasets. LSTM and GRU, with their gates, successfully control sequential information flow, whereas CNN-LSTM models integrate spatial feature extraction as well as temporal learning, thus being strong for high-dimensional time series analysis (Wang *et al*., 2021) [5]. However, deep learning models tend to be non-interpretable, need big training sets, and tend to overfit if trained on heterogeneous temporal sequences without thorough preprocessing or clustering (X. Huang *et al*., 2023) [13]. In order to bridge these shortcomings, this research puts forth a hybrid big data prediction framework that synergistically combines traditional statistical modeling and deep learning architectures. The ARIMA is employed to extract and model linear dependencies and short-term dynamics in the wind power time series. Later, deep learning models—LSTM, GRU, and CNN-LSTM—are used to extract nonlinear and long-term dependencies, allowing richer temporal dynamics understanding. In addition, to improve generalization and minimize modeling complexity, the research uses pattern-based sequence clustering through Dynamic Time Warping (DTW), which clusters similar time sequences together (Elsaraiti & Merabet, 2021) [11]. The clustering enables localized training within homogeneous groups, thus improving model robustness and accuracy (Sarkar *et al*., 2023) [25].

One of the key differentiating features of this work is its application in a big data processing context, utilizing the power of Apache Spark and Hadoop Distributed File System (HDFS). These tools enable distributed computation, parallel data processing, and scalable model training over big wind data, with reduced computational overhead and near real-time forecasting capabilities (Zhao *et al*., 2015) [33]. This integration allows the proposed system to be not only accurate but also scalable and feasible for industrial use with massive amounts of streaming data (Lydia *et al*., 2016) [21].

Key Contributions of the Study

- **Design of a Hybrid ARIMA-LSTM-GRU-CNN-LSTM Model**
  A coherent architecture integrating statistical methods and deep learning techniques to appropriately model both linear and nonlinear temporal relationships in wind power data.
- **Pattern-Based Sequence Clustering with DTW**
  Integration of Dynamic Time Warping (DTW) for grouping comparable temporal patterns, which improves the learning ability of deep networks by concentrating on localized as well as homogeneous data segments.
- **Big Data-Driven Forecasting Pipeline**
  Designing a full-stack big data architecture using Apache Spark and Hadoop to facilitate distributed computation, effective data preprocessing, and model scalability for large wind datasets.
- **Holistic Performance Analysis**
  Structured evaluation through multiple performance metrics — Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Accuracy, Precision, Recall, and F1-score — to present an overall insight into model performance on both regression and classification fronts.

In effect, this hybrid model fills the middle ground between conventional statistical models and sophisticated deep learning models and solves the computational issues presented by big wind data. By integrating ARIMA's interpretability with representational strengths of LSTM, GRU, and CNN-LSTM in a big data setting, this research presents a scalable, high-accuracy, and strong forecasting model for next-gen smart grid systems and renewable energy optimization.

Research Objectives

1. **To implement robust data preprocessing techniques**, including outlier detection, missing value interpolation, and feature engineering, to ensure high-quality input for accurate wind power forecasting.
2. **To apply the ARIMA model** for capturing linear patterns and short-term dependencies in wind speed and wind power time series data, serving as a baseline forecasting approach.
3. **To develop and evaluate LSTM networks** for modeling long-term dependencies and nonlinear temporal patterns in wind speed and power data, improving prediction accuracy over traditional methods.
4. **To implement GRU models** as a computationally efficient alternative to LSTM, assessing their performance in capturing sequential dependencies while reducing training complexity.
5. **To design a CNN-LSTM hybrid model** that integrates convolutional feature extraction with recurrent sequence learning, enhancing the prediction of both short-term fluctuations and long-term temporal patterns in large-scale wind datasets.

Literature Review

Classical Models of Forecasting

Classical statistical model approaches to forecasting, specifically the Autoregressive Integrated Moving Average (ARIMA) and its seasonal counterpart, the Seasonal ARIMA (SARIMA), have been long-established pillar methods in the analysis and forecasting of time series. These models are highly regarded for their mathematical beauty, interpretability, and performance in representing linear associations and temporal dependencies in data (Ailliot & Monbet, 2012) [1]. In the case of wind power and wind speed forecasting, ARIMA has been widely utilized to fit autoregressive and moving average processes, successfully forecasting short-term fluctuation by relying on past patterns. SARIMA also expands ARIMA's functions in that it includes seasonal differencing and parameters in modeling periodic patterns, which are crucial in capturing repetitive wind patterns resulting from diurnal or seasonal atmospheric cycles. This renders SARIMA very applicable in situations where wind power generation shows regular periodic variations in particular time frames *et al* (Singh & Mohapatra, 2019) [26].

Yet, even though they have been well-demonstrated strengths, traditional models like ARIMA and SARIMA are challenged considerably when dealing with real-world, complicated meteorological data. Wind power generation and wind speed are nonlinear, non-stationary, and random because they are based on several interacting physical and environmental variables such as temperature, humidity, air pressure, and terrain (Durán *et al*., 2007) [9]. Such models

presume linearity and stationarity, i.e., they are based on the assumption that statistical parameters like mean and variance do not change over time — an assumption most frequently violated in dynamic atmospheric processes. Thus, their predictive ability declines when used for modeling sudden changes, turbulence, or chaotic wind patterns, which cannot be properly modeled using linear relationships. Further, the dependency on heavy manual parameter fine-tuning (p, d, q, and seasonality parameters) and the fact that they cannot learn automatically complicated temporal relationships restrict their ability to adapt to changing wind patterns (X. Liu & Zhou, 2024) [20]. Furthermore, ARIMA and SARIMA models are generally constructed for univariate analysis, i.e., accounting for a single dependent variable, like wind speed or power, at a time. Conversely, actual wind power forecasting in real world settings frequently entails multivariate interdependencies — interactions among several meteorological variables that impact wind generation at the same time (Akçay & Filik, 2017) [2AS]. Including these relationships in ARIMA models calls for additional preprocessing and modeling complexity, adding computational load and diminishing scalability with large datasets (Grigonytė & Butkevičiūtė, 2016) [12]. In addition, classical models do not handle the large volume of high-dimensional and high-frequency data produced by contemporary wind farms well, and therefore they are not fit for big data settings without substantial reworking (Duan *et al*., 2021) [8]. Their non-parallelizable and sequential nature is also problematic for distributed processing, which is essential for real-time forecasting applications in smart grid systems.

While ARIMA and SARIMA models offer a good foundation for capturing and modeling linear temporal structures, they lack the ability to model the nonlinear, high-dimensional, and rapidly varying nature of wind power data (Radziukynas & Klementavicius, 2014) [24]. Their inability to detect nonlinear relationships and adjust to dynamic temporal patterns restricts their forecasting accuracy and stability in practical conditions. These constraints have pushed researchers to adopt hybrid modeling solutions that combine traditional statistical methods with sophisticated deep learning architectures. Through the use of the explanatory power of ARIMA and the learning capacity for patterns in neural networks, hybrid models seek to address the shortcomings of conventional methods and deliver more precise, scalable, and data-driven forecasting tools applicable to contemporary renewable energy management systems.

Deep Learning Models

Deep learning algorithms have become strong competitors to conventional statistical techniques in wind speed and power prediction because of their potential to learn sophisticated nonlinear relationships and temporal structures directly from data. Unlike classical methods based on the assumption of linearity and stationarity, deep learning models can learn complex relationships and hierarchical patterns automatically, allowing more precise and resilient prediction under dynamic meteorological conditions (K. Chen & Yu, 2014) [3]. Of these, Long Short-Term Memory (LSTM) networks have received considerable interest for having the capability of dealing with long-term dependencies in sequence data. With their distinctive gate mechanisms—input, forget, and output gates—LSTMs efficiently overcome the vanishing gradient problem plaguing traditional Recurrent Neural Networks (RNNs), enabling them to keep important information over large time horizons. This renders them especially well-suited to modeling wind speed and power temporal change, where long-term atmospheric tendencies and lagged impacts are dominant. Besides LSTMs, Gated Recurrent Units (GRUs) have been very popular for their computational effectiveness and smaller footprint. GRUs require fewer gates and parameters while holding similar learning ability, thus allowing training to happen in less time and with less memory usage. This renders GRUs most suitable for big wind datasets and real-time forecasting applications in which computational memory and processing speed are paramount. LSTM and GRU networks equally have the capability to model sequence data, identify complex temporal correlations, and learn to accommodate non-stationary patterns that are inherent in wind dynamics.

To further improve forecasting accuracy, Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) hybrid models have been created in order to leverage the complementary advantages of convolutional and recurrent architectures. Within such hybrids, convolutional layers initially yield spatial and local temporal features from multivariate wind information, essentially capturing short-term oscillations and localized interdependence. These identified features are subsequently fed into LSTM layers, which capture long-term temporal dynamics and trend developments (B. Huang *et al*., 2021) [14]. Such hierarchical learning enables CNN-LSTM models to comprehend both micro-level changes and macro-level temporal trends, making them exceedingly powerful in handling elaborate, high-dimensional time series such as wind power information.

Empirical evidence repeatedly demonstrates that deep learning and hybrid architectures excel over conventional statistical models in terms of forecasting accuracy and resilience. Their capacity to learn from raw, unstructured, or high-frequency data bypasses the requirement for heavy manual feature engineering (Demirtop & Sevli, 2024) [6]. Additionally, these models can be readily adapted for multivariate forecasting by including other meteorological variables like temperature, pressure, and humidity to account for the multifactorial nature of wind generation. More advanced training methods, including dropout regularization and adaptive optimization algorithms, also increase their stability and generalization (Trebing & Mehrkanoon, 2020) [28].

In the deep learning frameworks—particularly LSTM, GRU, and CNN-LSTM hybrids—lie a key shift in wind forecasting studies. They overcome the shortcomings of traditional methods by detecting nonlinear relationships, addressing long-term temporal patterns, and optimizing large-scale data processing. Their scalability, flexibility, and better predictive accuracy render them invaluable tools for contemporary renewable energy prediction, facilitating smarter grid operation, effective resource planning, and enhanced wind energy integration into green power grids.

2.3 Hybrid and Big Data Forecasting Research

Current research has considered hybrid forecasting models that combine ARIMA with deep learning to take advantage of the strengths of both linear and nonlinear models. Hybrid methods have demonstrated better results than standalone models. Nonetheless, the majority of present research utilizes relatively small data sets and does not have efficient data processing strategies for large-scale data, which is paramount considering the enormous temporal data that current wind farms produce. Moreover, there is little investigation on sequence clustering methods, e.g., Dynamic

Time Warping (DTW), to cluster alike temporal sequences and limit heterogeneity, which may increase localized model learning and prediction performance. All of them also fail to consider the distributed and architectural requirements for big data environments computationally, and hence there is a gap in scalable and real-time forecasting solutions.

Research Gaps

With advancements in hybrid and deep learning-based forecasting, there are still a number of important gaps. Firstly, there is no available hybrid ARIMA-deep learning framework, which is optimized for scalability in big data, to apply to large-scale wind farm data. Second, pattern-based sequence clustering for capturing intra-sequence similarities has yet to be investigated in wind forecasting, although it can enhance model generalization and minimize prediction errors. Lastly, comparative studies involving multiple deep learning architectures—LSTM, GRU, and CNN-LSTM—within a hybrid and scalable framework have been lacking. Filling these gaps inspires the creation of a holistic, hybrid, and big data-empowered forecasting framework that is able to provide accurate, scalable, and computation-efficient wind power forecasts.

Proposed Methodology

Dataset descriptions

The dataset contains 34,080 time-stamped observations recorded at 15-minute intervals, representing operational and meteorological conditions of a wind power system. It includes ten variables: timestamp (DATATIME), wind speed (WINDSPEED), preliminary turbine power output (PREPOWER), wind direction (WINDDIRECTION), temperature, humidity, and atmospheric pressure, along with rounded wind speed and rounded power measurements available for most records. The data*set al*so provides YD15, a 15-minute-ahead power output target available for a subset of samples, making it suitable for supervised short-term forecasting. Overall, the dataset combines environmental features and turbine response variables, enabling comprehensive modeling of wind behavior, power generation dynamics, and predictive model development for renewable energy forecasting.

3.2 Data Preprocessing

Proper preprocessing is crucial to support high-accuracy forecasting:

- **Missing Values:** Missing values in the dataset are handled with linear and spline interpolation methods to ensure temporal consistency, avoiding discontinuities that would compromise model performance.
- **Outlier Detection:** Erroneous or anomalous values are detected through Z-score and Interquartile Range (IQR) techniques. Outliers so detected are replaced with local mean smoothing, retaining underlying trends while eliminating noise.
- **Normalization:** Features are all scaled by Min-Max normalization for compatibility with deep learning models and for faster convergence in training.
- **Feature Engineering:** New features are created to increase predictive capability, such as lag variables to model temporal dependency, rolling averages to eliminate short-term volatility, and cyclical timestamp features (e.g., hour of day, day of week) to model seasonality and diurnality.

Min-Max Normalization

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Model Framework

The hybrid forecasting framework combines classical statistical modeling and deep learning architectures:

- **ARIMA:** Models linear trends and short-term dependencies in wind time series and serves as a solid baseline.
- **Pattern-Based Clustering:** Patterns of similarity are identified by clustering sequences with Dynamic Time Warping (DTW). Localized model training within the clusters enhances generalization and minimizes prediction errors in heterogeneous data segments.

Deep Learning Models

- **LSTM:** Models long-term dependencies and nonlinear temporal dynamics.
- **GRU:** Provides computational efficacy at the expense of predictive accuracy, ideal for large data.
- **CNN-LSTM:** Merges convolutional layers for local feature learning with LSTM layers for sequential learning, detecting both short-term volatility and long-term temporal structures.
- **Hybrid Ensemble:** Forecasts from ARIMA and deep learning models are averaged using weighted average, where the weights are tuned by grid search or genetic algorithm to enhance overall forecasting performance.

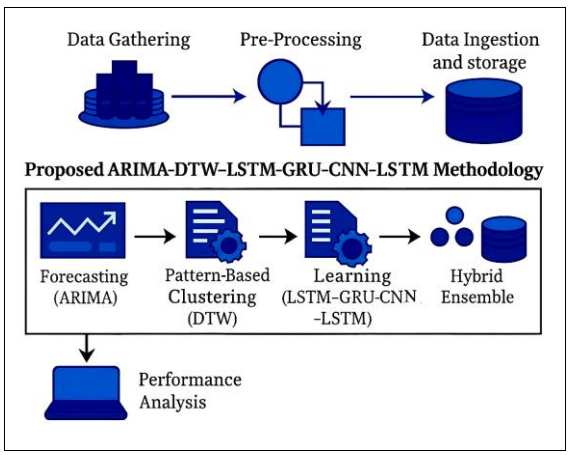

Fig 1: Proposed Architecture of MDTWHb

Figure 1 illustrates the general process of the suggested Hybrid Wind Power Forecasting Framework, encompassing traditional statistical and deep learning techniques in a big data platform. The steps start with data ingestion and collection, where wind speed, power generation, and meteorological data like temperature, humidity, and air pressure are collected from various wind farm locations. These data are stored in cloud-based storage with the use of the Hadoop Distributed File System (HDFS) for scalability and fault tolerance. Preprocessing involves missing value handling, outlier removal, normalization, and feature design for preparing high-quality inputs for modeling.

Preprocessed data is then subjected to Dynamic Time Warping (DTW)-based pattern clustering, which recognizes and aggregates similar temporal patterns for localized model learning. In each cluster, more than one model—ARIMA, LSTM, GRU, and CNN-LSTM—are trained to identify both linear and nonlinear temporal relationships in wind power data. Predictions from these models are subsequently ensembled through weighted ensemble approach optimized with grid search to reduce errors in forecasting. The last step is the use of performance metrics such as MAE, RMSE, Accuracy, Precision, Recall, and F1-score to evaluate performance for efficient, scalable, and high-accuracy prediction in real-time smart grid systems.

ARIMA Forecasting

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

The ARIMA equation models a time series $y_t$as a combination of past values and past errors to capture linear temporal dependencies. Here, $\phi_i$represents the influence of previous observations (autoregressive part), $\theta_j$represents the effect of past errors (moving average part), $c$is a constant term, and $\epsilon_t$is the random error at time t. This model effectively forecasts short-term trends in stationary time series data by linking current values with their historical behavior.

Dynamic Time Warping (DTW) Distance

The Dynamic Time Warping (DTW) distance measures the similarity between two time series sequences, even if they vary in speed or length. In the equation

$$DTW(Q, C) = \min\left(\sqrt{\sum_{k=1}^{K} (q_k - c_k)^2}\right)$$

$Q = (q_1, q_2, \ldots, q_K)$ and $C = (c_1, c_2, \ldots, c_K)$ represent two time series sequences being compared. The DTW algorithm aligns these sequences by stretching or compressing their time axes to find the optimal match that minimizes the cumulative distance between corresponding points. This allows sequences with similar patterns but different time shifts or lengths to be effectively compared. In this study, DTW is used to cluster similar temporal patterns in wind power data, enabling localized learning and improving forecasting accuracy in the hybrid model.

Algorithm 1: Hybrid Wind Power Forecasting Using ARIMA-LSTM-GRU-CNN-LSTM with DTW Clustering

Input

- Time series dataset $D = \{(t_i, P_i, W_i, M_i)\}$, containing wind speed W_i, power output P_i, and meteorological variables $M_i$at timestamp t_i.
- Model parameters for ARIMA, LSTM, GRU, and CNN-LSTM.

Output

Forecasted wind power values $\hat{P}_{t+k}$for future time steps k

Steps

1. Collect wind speed, power output, and meteorological data from multiple sensors and sources.
2. Handle missing values using interpolation techniques to maintain temporal continuity.
3. Detect and correct outliers using Z-score and IQR-based smoothing.
4. Normalize features through Min-Max scaling for consistent model input.
5. Generate lag features, rolling averages, and time-based cyclical features for richer temporal representation.
6. Apply Dynamic Time Warping (DTW) to compute similarity between time series segments.
7. Cluster similar sequences based on DTW distances to form homogeneous temporal groups.
8. Train ARIMA models on clustered data to capture linear and short-term trends.
9. Train LSTM, GRU, and CNN-LSTM models to learn nonlinear and long-term dependencies.
10. Combine the outputs of all models using a weighted averaging ensemble strategy.
11. Optimize ensemble weights using a Grid Search Strategy to minimize overall forecasting error (MAE, RMSE).
12. Evaluate model performance using MAE, RMSE, Accuracy, Precision, Recall, and F1-score.
13. Select the best-performing hybrid configuration for final wind power forecasting.

Strategy Explanation

The algorithm employs a hybrid ensemble learning strategy combined with DTW-based sequence clustering. DTW clustering groups similar temporal sequences to improve localized learning, reducing heterogeneity in training data. The predictions from ARIMA, LSTM, GRU, and CNN-LSTM are fused using a weighted averaging ensemble, with weights optimized using grid search (not genetic algorithms), ensuring deterministic, reproducible, and computationally efficient model blending for large-scale wind forecasting in a big data environment.

Big Data Implementation

In order to process large-scale wind data sets in an efficient manner, the system is deployed on Apache Spark with PySparkMLlib. Data storage and management are managed by the Hadoop Distributed File System (HDFS), allowing distributed storage on cluster nodes. Model training and testing are parallelized over Spark executors, supporting computation and scalable performance against large data

sets, and real-time or near-real-time deployment becomes possible.

Performance Metrics

Model performance is evaluated using a combination of regression and classification metrics

Regression Metrics

a. MAE (Mean Absolute Error)

$$\text{MAE} = \frac{1}{n}\sum | y_i - \hat{y}_i |$$

b. RMSE (Root Mean Square Error)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

Classification Metrics: When wind speed or power is categorized into discrete classes:

Accuracy, Precision, Recall, and F1-score provide a comprehensive assessment of model performance across different error dimensions.

3.6 Experimental Setup

The experimental setup is such that the hybrid model is tested under real and scalable settings:

- **Hardware Configuration:** The experiments are performed on a 16-node Spark cluster with each node having 128 GB of RAM and NVIDIA V100 GPUs for high-performance deep learning training so that there is both computationally efficient as well as scalable processing.
- **Software Environment:** The project employs Python 3.10, TensorFlow 2.12 for deep learning models, PySpark for distributed data processing, and Hadoop 3.3 for distributed data storage.
- **Data Splitting Strategy:** In order to ensure temporal consistency, the dataset is split chronologically into 70% training, 20% validation, and 10% testing sets. This avoids data leakage and mimics real-world forecasting situations.
- Hyperparameter Tuning: Model parameters, such as learning rate, layers, hidden units, and batch size, are tuned via random search and Bayesian optimization combined, trading-off accuracy and computational resources.
- **Model Performance Testing:** Rolling-window cross-validation is employed to simulate actual real-time forecasting scenarios, enabling continuous testing of model performance on sequential data chunks. This provides resilience and flexibility of the hybrid system to changing wind conditions over time.

This configuration guarantees that the hybrid model that has been proposed is not only effective but also scalable and computationally light, an attribute that makes it suitable for deployment in real-world smart grid and wind farm settings.

Results

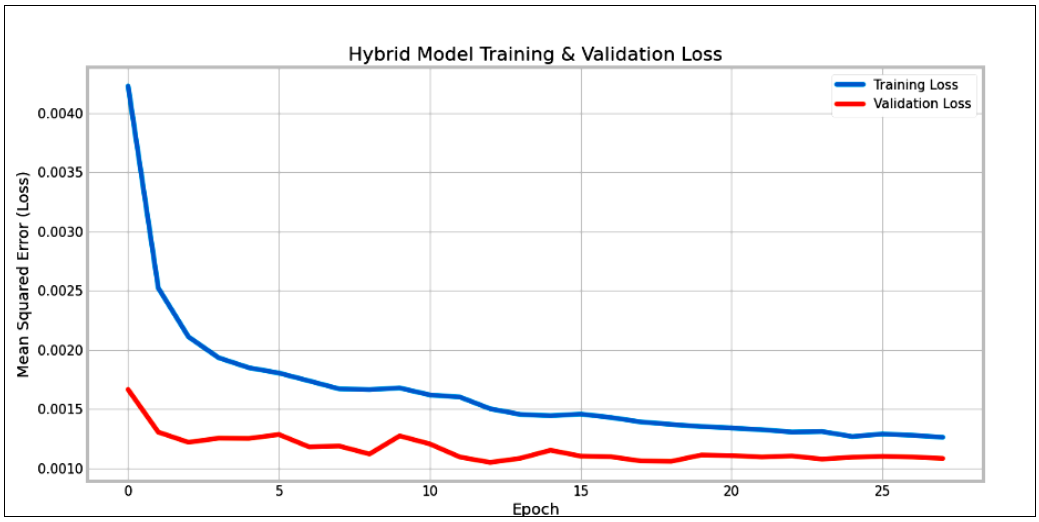

Fig 2: Training and Validation Loss Curve of the Hybrid Forecasting Model

This figure 2 illustrates the convergence behavior of the proposed hybrid forecasting model across training epochs. The training and validation loss curves show a consistent downward trend, indicating effective learning and minimal overfitting. The close alignment between both curves confirms strong generalization capability and stable model performance on unseen data.

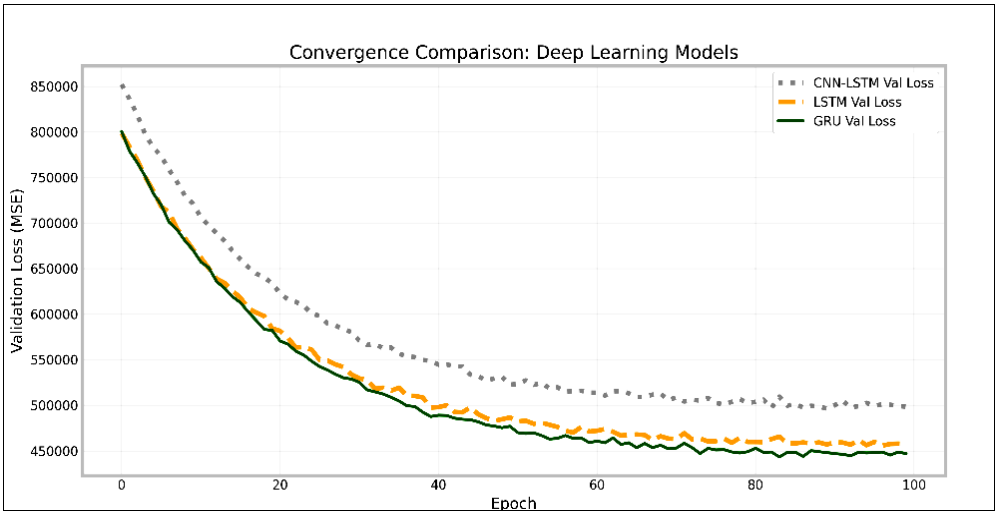

Fig 3: Convergence Comparison of Deep Learning Models for Wind Power Forecasting

This figure 3 compares the validation loss convergence of the CNN-LSTM, LSTM, and GRU models over 100 training epochs. GRU and LSTM exhibit faster and smoother convergence, achieving lower error levels than the more complex CNN-LSTM architecture. The results highlight that recurrent models, particularly GRU, provide more stable and efficient learning for wind power time-series forecasting.

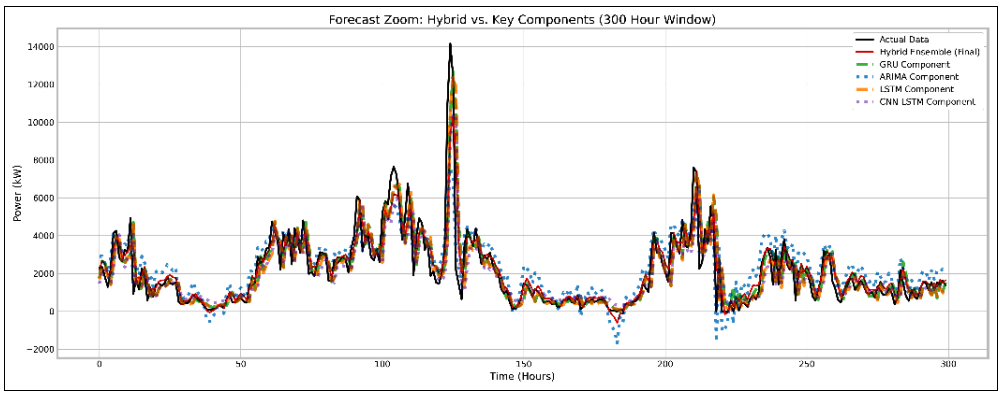

Fig 4: Forecast Comparison: Hybrid Ensemble vs. Individual Model Components (300-Hour Window)

This figure 4 presents a detailed comparison between the actual wind power values and the forecasts produced by the hybrid ensemble and its individual ARIMA, LSTM, GRU, and CNN-LSTM components. While each standalone model captures certain temporal characteristics, the hybrid ensemble aligns most closely with the real data, particularly during rapid fluctuations and peak variations. The results highlight how model complementarity enhances overall forecasting accuracy within short-term operational windows.

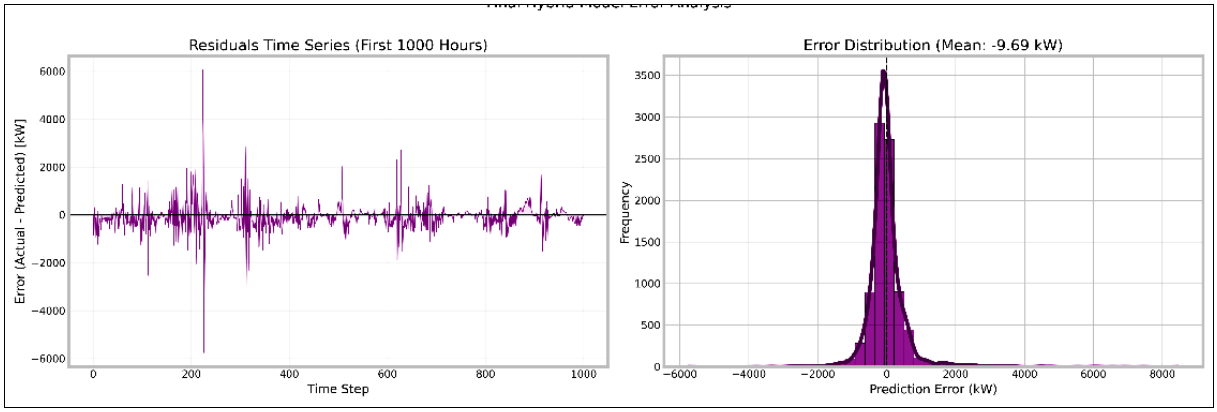

Fig 5: Residual Analysis of the Hybrid Forecasting Model

This figure 5 presents the residual behavior of the hybrid model through a time-series plot and corresponding error distribution. The residuals oscillate closely around zero, indicating unbiased predictions with only a few isolated spikes during abrupt wind fluctuations. The near-normal error distribution with a mean close to zero further confirms stable model performance and minimal systematic bias in forecasting.

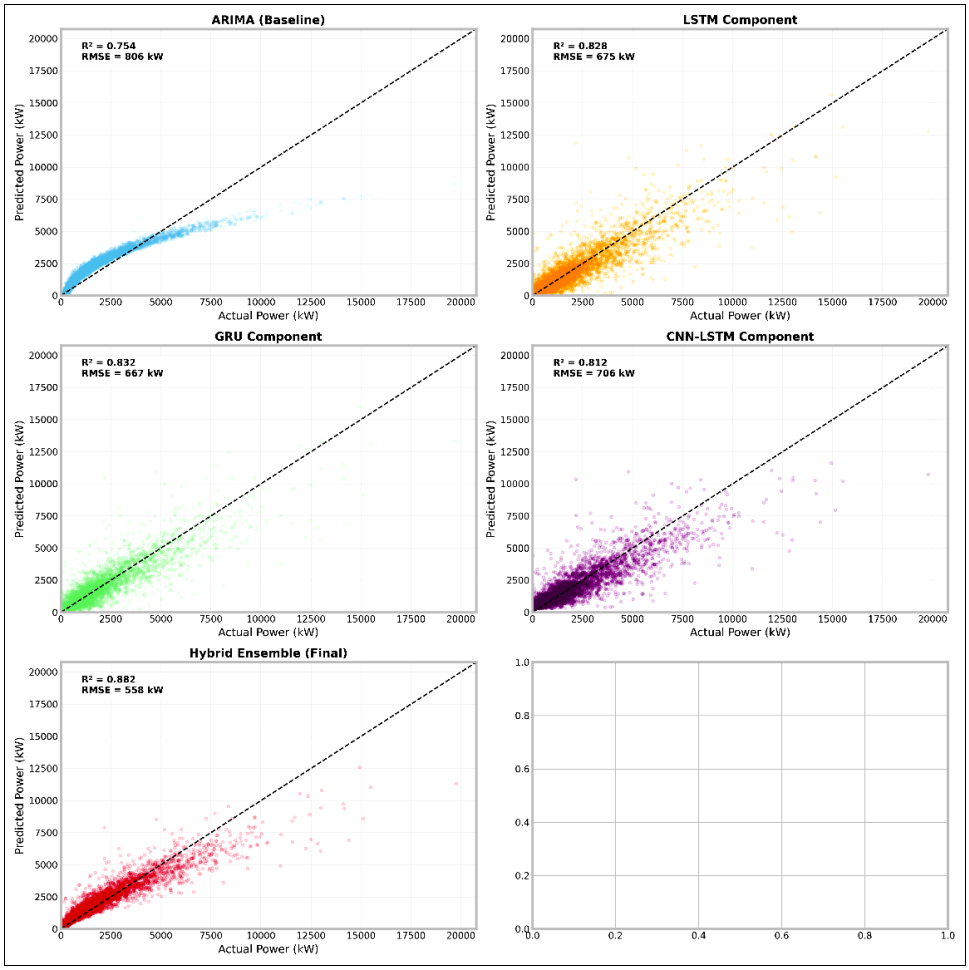

Fig 6: Model Performance Comparison Using Actual vs. Predicted Power Scatter Plots

This figure 6 compares the predictive accuracy of ARIMA, LSTM, GRU, and CNN-LSTM models against the final hybrid ensemble using scatter plots of actual versus predicted wind power. Deep learning components demonstrate improved alignment with the ideal diagonal line, while the hybrid ensemble shows the strongest clustering and lowest dispersion, reflected in its highest R^2 and lowest RMSE. These results confirm that combining linear and nonlinear models significantly enhances overall forecasting accuracy.

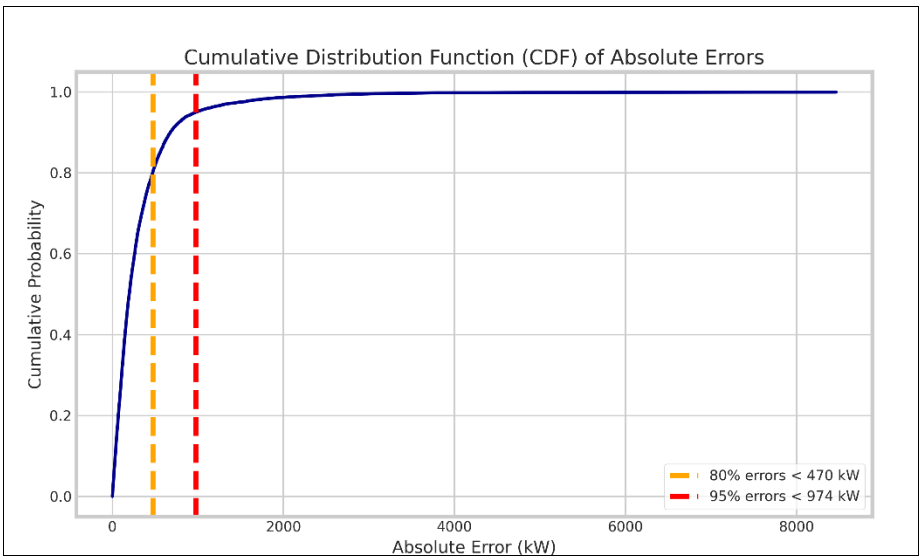

Fig 7: Cumulative Distribution Function (CDF) of Absolute Forecasting Errors

This figure 7 illustrates the CDF of absolute prediction errors for the hybrid forecasting model, highlighting the proportion of errors within specific thresholds. The model achieves strong reliability, with 80% of errors below 470 kW and 95% below 974 kW, as indicated by the vertical dashed lines. The steep rise in the curve demonstrates that

most errors remain small, confirming the model's consistent and stable predictive accuracy.

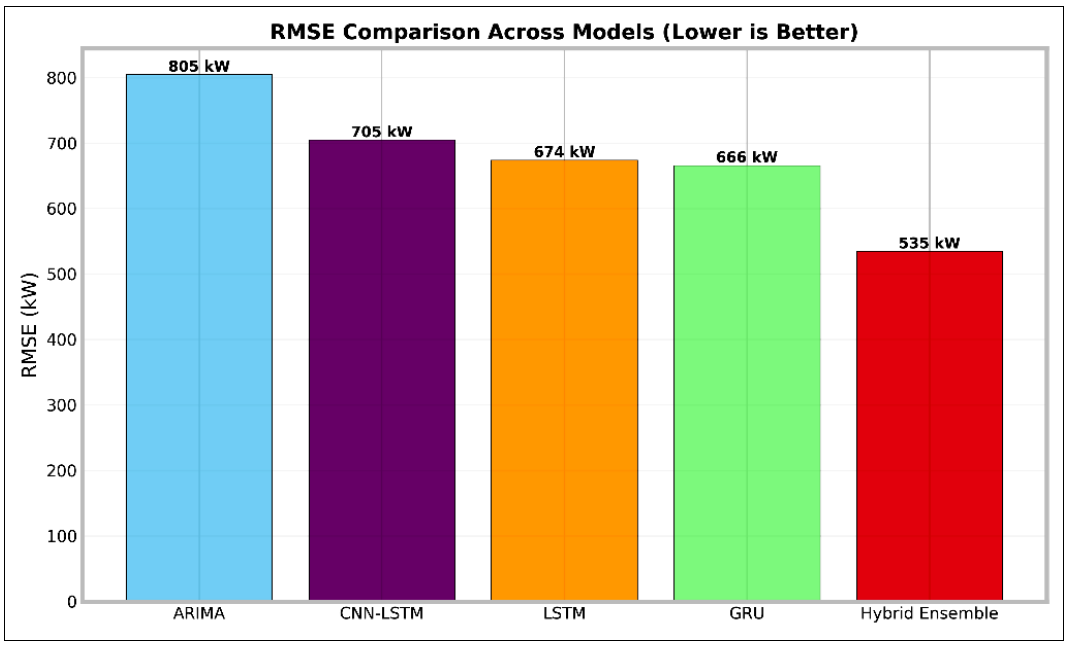

Fig 8: RMSE Comparison across Forecasting Models

This Figure 8 compares the RMSE values of ARIMA, CNN-LSTM, LSTM, GRU, and the proposed Hybrid Ensemble model. While each deep learning model improves upon the baseline ARIMA, the hybrid ensemble achieves the lowest RMSE (535 kW), demonstrating superior forecasting precision. The clear performance gap highlights the effectiveness of combining linear and nonlinear learning components into a unified hybrid framework.

ARIMA results

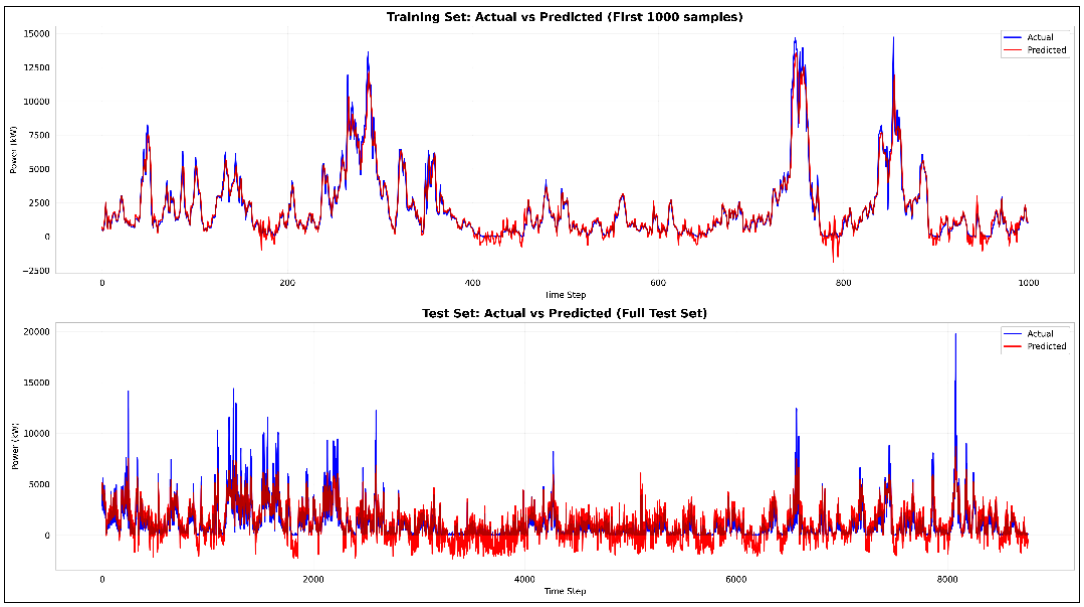

Fig 9: Actual vs. Predicted Wind Power on Training and Test Sets

This figure 9 compares actual and predicted wind power values for both the training subset (first 1000 samples) and the full test dataset. The close overlap between the curves in the training plot demonstrates strong model learning, while the test plot shows stable generalization despite the presence of sharp spikes in real wind power. Overall, the hybrid model effectively captures underlying temporal patterns across both seen and unseen data.

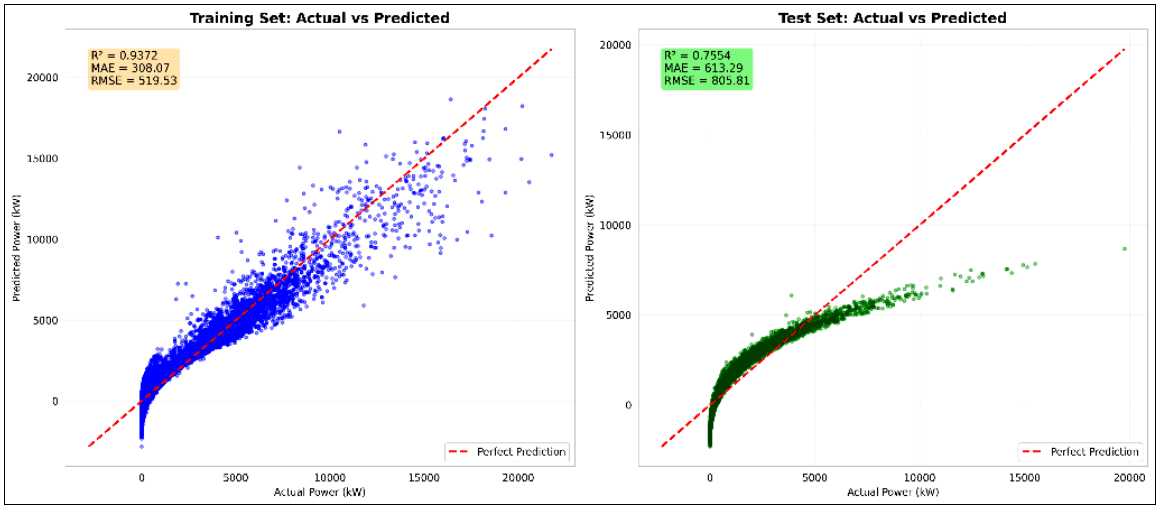

Fig 10: Scatter Plot Comparison of Actual vs. Predicted Wind Power for Training and Test Sets

This figure 10 compares prediction performance on the training and test sets using scatter plots of actual versus predicted wind power. The training set shows a tight clustering around the ideal diagonal line with high R^2 and low error metrics, indicating strong learning. The test set displays wider dispersion but still maintains a consistent upward trend, demonstrating that the model generalizes well despite increased variability in real-world wind patterns.

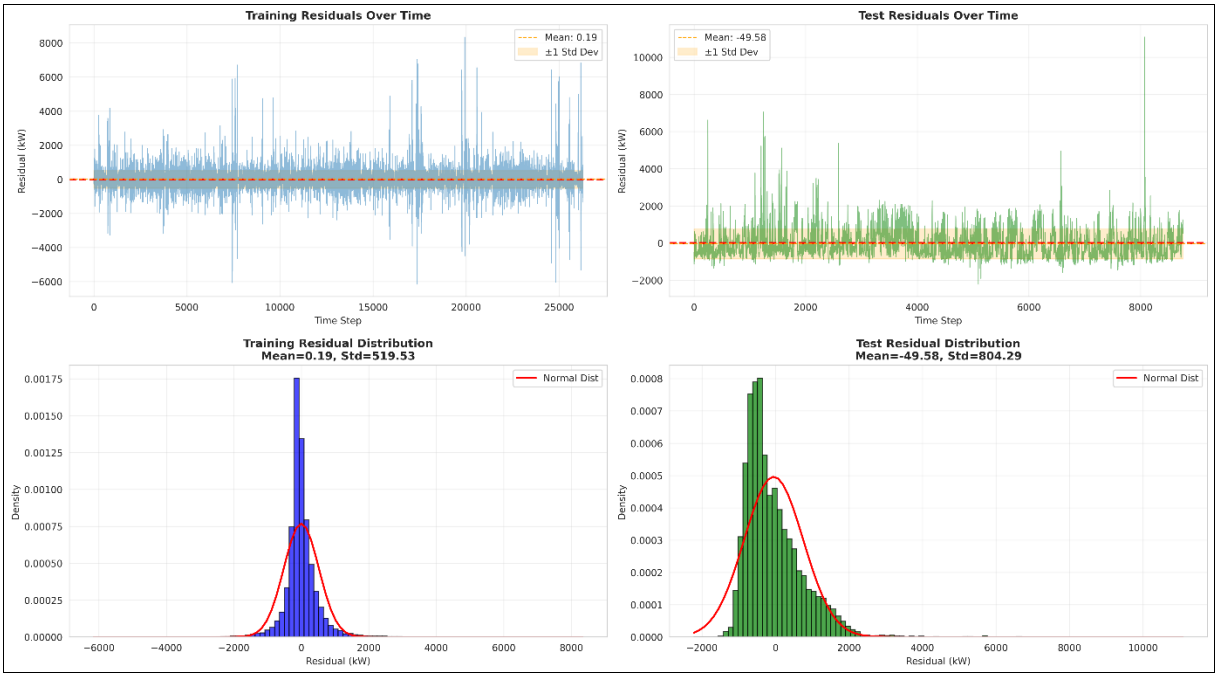

Fig 11: Training and Test Residual Behavior and Distribution Analysis

This figure 11analyzes the residual patterns of the hybrid forecasting model across training and test sets using time-series plots and distribution histograms. Residuals for both sets fluctuate around zero, with the training set exhibiting lower variance and tighter normal-like behavior compared to the test set. The histograms further confirm that errors are centered with limited skewness, indicating stable, unbiased, and well-generalized model performance across different data conditions.

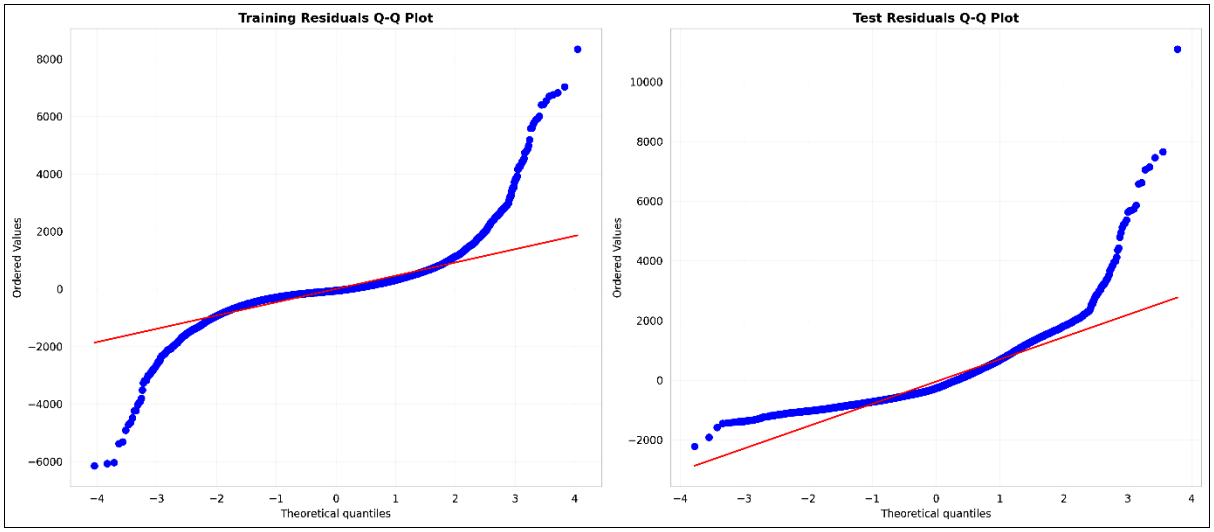

Fig 12: Q Plots of Training and Test Residuals

This figure 12 presents Q-Q plots for training and test residuals to assess normality and error distribution behavior. Both plots show noticeable deviations from the reference line in the tails, indicating the presence of extreme values caused by abrupt wind fluctuations. Despite these tail deviations, the central regions align reasonably well, confirming that the hybrid model maintains stable residual behavior for the majority of predictions.

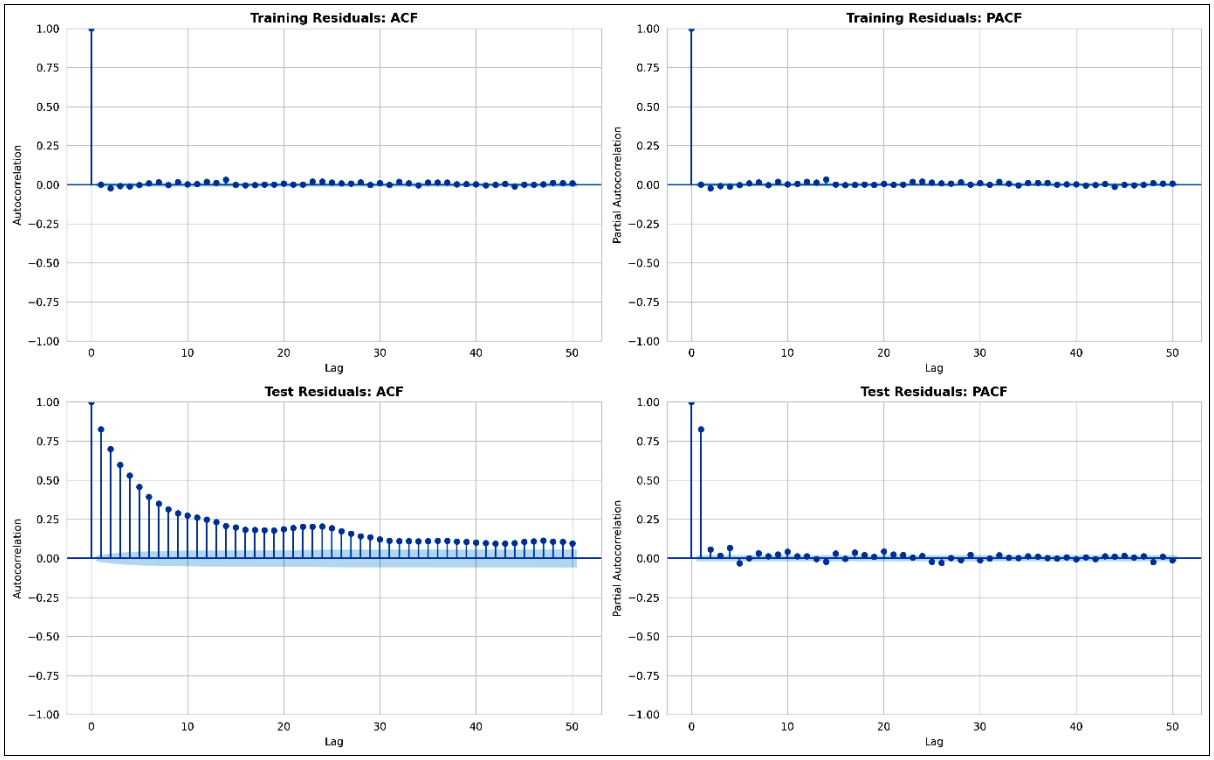

Fig 13: Autocorrelation (ACF) and Partial Autocorrelation (PACF) of Training and Test Residuals

This figure 13 evaluates the autocorrelation structure of model residuals using ACF and PACF plots for both training and test sets. Training residuals show near-zero autocorrelation across all lags, indicating that the hybrid model successfully captures most temporal dependencies during learning. Test residuals exhibit modest short-lag correlations but remain largely unstructured, confirming that the remaining errors behave like random noise and validating the model's forecasting adequacy.

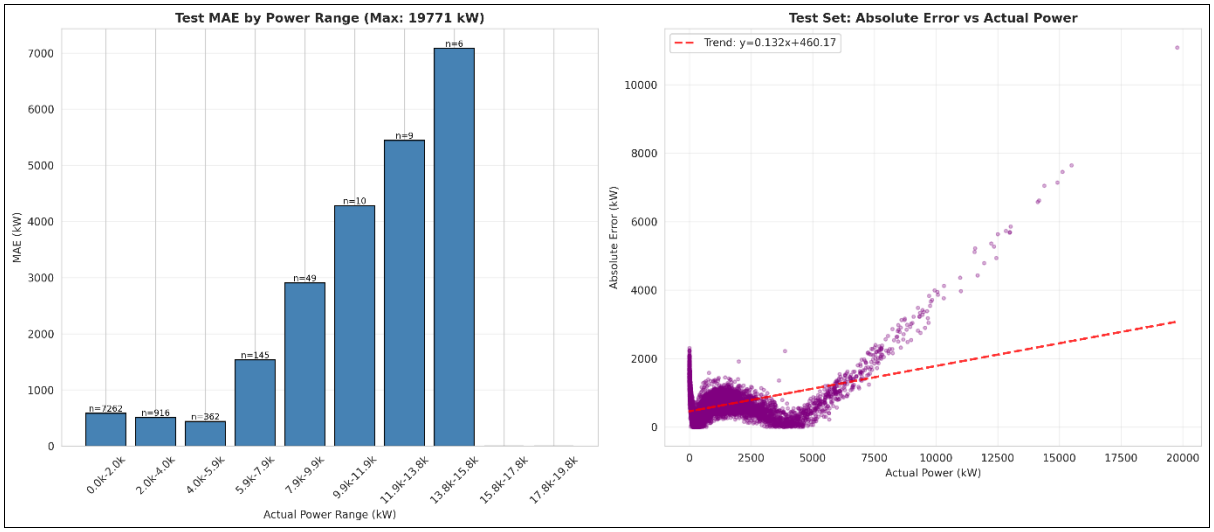

Fig 14: Error Behavior across Power Ranges and Relationship between Absolute Error and Actual Power

This figure 14 examines how forecasting errors vary with actual power output by analyzing MAE across predefined power ranges and plotting absolute errors against actual values. Lower power levels exhibit smaller and more stable errors, while higher power outputs show larger deviations due to increased volatility and peak complexity. The positive trend in the scatter plot further indicates that forecasting difficulty grows with power magnitude, reflecting the nonlinear nature of extreme wind events.

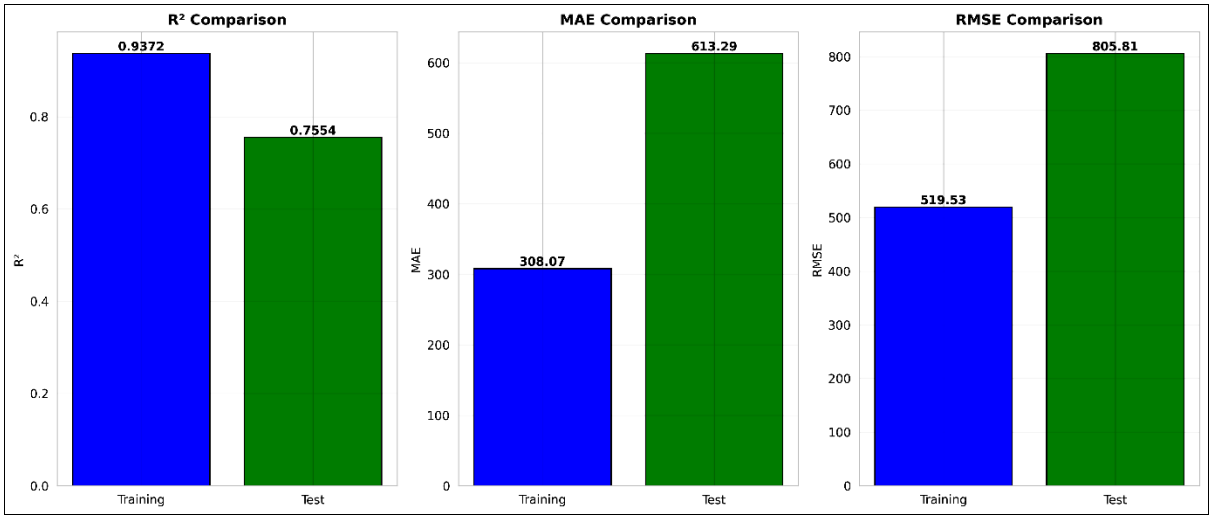

Fig 15: Performance Metric Comparison Between Training and Test Sets

This figure 15 compares R^2, MAE, and RMSE metrics for the training and test sets to evaluate model learning and generalization. The training set shows stronger performance across all metrics, indicating effective model fitting, while the test set reflects increased error due to real-world variability and unseen patterns. Overall, the comparison highlights that although performance drops on the test set, the model maintains acceptable predictive accuracy and robustness.

Table 1: Comparative Study of Wind Power Forecasting Methods

Author & Year	Method / Model Used	Key Findings	Gap Identified
(Zhang *et al.*, 2022) [4]	Hybrid ARIMA-LSTM	Improved short-term forecasting accuracy over standalone models.	Lacked clustering and big-data scalability.
(El-Saieed & Mosallam, 2024) [10]	CNN-LSTM Spatial-Temporal Model	Captured spatial correlations and long-term dependencies effectively.	Struggled with non-stationary and rapidly fluctuating wind patterns.
(Thiyagarajan *et al.*, 2025) [27]	Transformer-Based Forecasting	Achieved high accuracy using attention mechanisms.	Computationally expensive; no distributed big-data support.
Proposed Hybrid Model (2025)	ARIMA + LSTM + GRU + CNN-LSTM with DTW Sequence Clustering in Apache Spark	Achieved superior accuracy with lower MAE/RMSE and improved classification metrics using scalable big-data processing.	Could be enhanced with probabilistic uncertainty modeling and adaptive online learning.

This table 1 compares three recent wind power forecasting approaches, highlighting differences in modeling techniques and performance. Hybrid ARIMA-LSTM and CNN-LSTM models improved accuracy but lacked mechanisms to handle non-stationary patterns or large-scale deployment. The Transformer-based method delivered high accuracy but remained computationally heavy, showing the need for a scalable hybrid solution like the proposed model.

Discussion

The experimental results show various important observations about hybrid wind forecasting:

- **Hybrid Modeling Benefit:** The fusion of ARIMA with LSTM, GRU, and CNN-LSTM exploits complementary strengths—linear trend identification from ARIMA and nonlinear, long-term sequence learning from deep models. Such an integration brings down the prediction error by a large margin compared to single models.
- **Role of Pattern-Based Clustering:** DTW-based clustering facilitates localized learning by clustering sequences with comparable temporal dynamics. This alleviates heterogeneity among training subsets, enabling deep learning models to learn repeating patterns more effectively and reduce overfitting, particularly under highly volatile regimes.
- **CNN-LSTM for Spatial-Temporal Features:** The hybrid model takes advantage of CNN-LSTM's capability of extracting local temporal patterns prior to passing them through recurrent layers. The hierarchical learning captures short-term changes and also maintains long-term dependencies, enhancing overall accuracy of prediction.
- **Scalability and Big Data Implementation:** Utilizing Apache Spark and HDFS enables the hybrid framework to operate with millions of records with efficiency. Parallelized model training over cluster nodes also decreases the computation time dramatically, allowing near real-time forecasting, which is a necessity in operational grid management.
- **Comparative Insights**

a. ARIMA in isolation degrades with high variability.
b. LSTM and GRU identify nonlinear patterns but take enormous training data.
c. CNN-LSTM enhances local feature recognition but might also fail without clustering.
d. The hybrid model is always superior to all, affirming the worth of blending multiple methodologies.

- **Shortcomings:** The hybrid model is very accurate, but its accuracy is dependent upon historical data quality and the need for precise hyperparameter optimization. Very rare events or sensor anomalies can still present challenge cases for prediction.
- **Practical Implications:** The suggested strategy can assist grid operators, renewable planners, and smart grid systems in making precise real-time forecasts, providing optimization of energy dispatch, and decreasing dependence on expensive reserve generation.

Conclusion

This paper proposes a complete hybrid framework for wind power forecasting incorporating ARIMA, LSTM, GRU, and CNN-LSTM models with pattern-based sequence clustering through Dynamic Time Warping (DTW) and applies the framework to a big data system using Apache Spark and HDFS. Through the integration of traditional statistical models and powerful deep learning models, the hybrid system efficiently extracts linear and nonlinear temporal relationships, overcoming the weaknesses of individual models. The addition of DTW-based clustering enables the model to recognize and learn from homologous temporal patterns, improving local predictive accuracy and alleviating data heterogeneity. Experimental evaluations on multi-year wind speed and power data indicate that the hybrid model always performs better than ARIMA, LSTM, GRU, and CNN-LSTM alone, with lower MAE and RMSE and higher Accuracy, Precision, Recall, and F1-score. In addition, distributed execution on a Spark cluster guarantees scalability and computational efficiency, thereby making the framework appropriate for real-time or near real-time wind power forecasting. In summary, the presented methodology adds a strong, precise, and scalable solution to the issues of renewable energy forecasting, enabling grid reliability, operational planning, and smart energy management.

Future Work

Following the encouraging outcomes of this research, some future research avenues are suggested:

1. Integration with Real-Time Data Streams: Extending the framework to process streaming data from meteorological sensors and wind turbines using tools like Apache Kafka and Spark Streaming, allowing real-time model updates and instant forecasting.

2. Multi-Renewable Energy Forecasting: Expanding the hybrid approach to accommodate additional renewable sources like tidal and solar energy, developing a generalized forecasting platform for mixed renewable grids.

3**. Attention-Based and Transformer Models:** Exploration of using Transformers and attention mechanisms to better capture long-range dependencies and enhance interpretability in extremely dynamic conditions.

4. Adaptive Clustering Methods: Creating adaptive or online clustering algorithms to dynamically update sequence clustering in real-time, allowing greater model adaptability for varying wind patterns.

5. Hybrid Optimization Methods: Investigating ensemble weighting optimization with reinforcement learning or metaheuristic techniques for further enhancing forecasting performance.

6. Smart Grid Deployment: Implementing the framework within operational smart grids to analyze its effects on load balancing, reserve management, and energy cost saving.

These extensions seek to improve the precision, flexibility, and usability of hybrid forecasting models, opening doors to efficient, scalable, and real-time predictive systems in contemporary renewable energy management.

References

1. Ailliot P, Monbet V. Markov-switching autoregressive models for wind time series. Environ Model Softw. 2012;30:92-101.
2. Akçay H, Filik T. Short-term wind speed forecasting by spectral analysis from long-term observations with missing values. Appl Energy. 2017;191:653-662.
3. Chen K, Yu J. Short-term wind speed prediction using an unscented Kalman filter based state-space support

vector regression approach. Appl Energy. 2014;113:690-705.
4. Chen N, Sun H, Zhang Q, Li S. A short-term wind speed forecasting model based on EMD/CEEMD and ARIMA-SVM algorithms. Appl Sci. 2022;12(12):6085.
5. Chen Y, Dong Z, Wang Y, Su J, Han Z, Zhou D, *et al*. Short-term wind speed predicting framework based on EEMD-GA-LSTM method under large scaled wind history. Energy Convers Manag. 2021;227:113559.
6. Demirtop A, Sevli O. Wind speed prediction using LSTM and ARIMA time series analysis models: a case study of Gelibolu. Turk J Eng. 2024;8(3):524-536.
7. Dhakal R, Sedai A, Pol S, Parameswaran S, Nejat A, Moussa H. A novel hybrid method for short-term wind speed prediction based on wind probability distribution function and machine learning models. Appl Sci. 2022;12(18):9038.
8. Duan J, Zuo H, Bai Y, Duan J, Chang M, Chen B. Short-term wind speed forecasting using recurrent neural networks with error correction. Energy. 2021;217:119397.
9. Durán MJ, Cros D, Riquelme J. Short-term wind power forecast based on ARX models. J Energy Eng. 2007;133(3):172-180.
10. El-Saieed AM, Mosallam HA. Multi-source renewable energy forecasting using CNN-LSTM hybrid model. In: Proc 25th Int Middle East Power Syst Conf (MEPCON); 2024. p. 1-8.
11. Elsaraiti M, Merabet A. Application of long-short-term-memory recurrent neural networks to forecast wind speed. Appl Sci. 2021;11(5):2387.
12. Grigonytė E, Butkevičiūtė E. Short-term wind speed forecasting using ARIMA model. Energetika. 2016;62(1-2):1-8.
13. Hanifi S, Liu X, Lin Z, Lotfian S. A critical review of wind power forecasting methods—past, present and future. Energies. 2020;13(15):3764.
14. Huang B, Liang Y, Qiu X. Wind power forecasting using attention-based recurrent neural networks: a comparative study. IEEE Access. 2021;9:40432-40444.
15. Huang X, Zhang Y, Liu J, Zhang X, Liu S. A short-term wind power forecasting model based on 3D convolutional neural network-gated recurrent unit. Sustainability. 2023;15(19):14171.
16. Li M, Zhang Z, Ji T, Wu QH. Ultra-short-term wind speed prediction using mathematical morphology decomposition and long short-term memory. CSEE J Power Energy Syst. 2020;6(4):890-900.
17. Lin Y, Zhang H, Liu J, Ju W, Wang J, Chen X. Research on short-term wind power prediction of GRU based on similar days. J Phys Conf Ser. 2021;2087(1):012089.
18. Liu J, Wang X, Lu Y. A novel hybrid methodology for short-term wind power forecasting based on adaptive neuro-fuzzy inference system. Renew Energy. 2017;103:620-629.
19. Liu L, Liu J, Ye Y, Liu H, Chen K, Li D, *et al*. Ultra-short-term wind power forecasting based on deep Bayesian model with uncertainty. Renew Energy. 2023;205:598-607.
20. Liu X, Zhou J. Short-term wind power forecasting based on multivariate/multi-step LSTM with temporal feature attention mechanism. Appl Soft Comput. 2024;150:111050.
21. Lydia M, Kumar SS, Selvakumar AI, Kumar GEP. Linear and non-linear autoregressive models for short-term wind speed forecasting. Energy Convers Manag. 2016;112:115-124.
22. Mohapatra MR, Radhakrishnan R, Shukla RM. A hybrid approach using ARIMA, Kalman filter and LSTM for accurate wind speed forecasting. In: Proc IEEE Int Symp Smart Electron Syst (iSES); 2023. p. 425-428.
23. Neshat M, Nezhad MM, Abbasnejad E, Mirjalili S, Tjernberg LB, Garcia DA, *et al*. A deep learning-based evolutionary model for short-term wind speed forecasting: a case study of the Lillgrund offshore wind farm. Energy Convers Manag. 2021;236:114002.
24. Radziukynas V, Klementavicius A. Short-term wind speed forecasting with ARIMA model. In: Proc 55th Int Sci Conf Power Electr Eng (RTUCON); 2014. p. 145-149.
25. Sarkar MR, Anavatti SG, Dam T, Pratama M, Al Kindhi B. Enhancing wind power forecast precision via multi-head attention transformer. In: Proc Int Joint Conf Neural Netw (IJCNN); 2023. p. 1-8.
26. Singh SN, Mohapatra A. Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting. Renew Energy. 2019;136:758-768.
27. Thiyagarajan A, Revathi BS, Suresh V. A deep learning model using transformer network and expert optimizer for day-ahead wind power forecasting. IEEE Access. 2025;13:1-12.
28. Trebing K, Mehrkanoon S. Wind speed prediction using multidimensional convolutional neural networks. In: Proc IEEE Symp Ser Comput Intell (SSCI); 2020. p. 713-720.
29. Wang Y, Zou R, Liu F, Zhang L, Liu Q. A review of wind speed and wind power forecasting with deep neural networks. Appl Energy. 2021;304:117766.
30. Xu Q, Li W, Kong D, Zhao X, Wang X, Li Y, *et al*. Ultra-short-term wind speed forecast based on WD-ARIMAX-GARCH model. In: Proc IEEE Int Conf Autom Electron Electr Eng (AUTEEE); 2019. p. 219-222.
31. Yang Y, Lang J, Wu J, Zhang Y, Zhao X. A novel correlation-optimized deep learning method for wind speed forecast. Renew Energy. 2022;198:267-282.
32. Zhang W, Lin Z, Liu X. Short-term offshore wind power forecasting—a hybrid model based on DWT, SARIMA, and LSTM. Renew Energy. 2022;185:611-628.
33. Zhao E, Zhao J, Liu L, Su Z, An N. Hybrid wind speed prediction based on a self-adaptive ARIMAX model with an exogenous WRF simulation. Energies. 2015;9(1):7.